ED 428 117 TM 029 588

AUTHOR	Brown, Diane Peacock
TITLE	Fragile--Handle with Care: Regression Analyses That Include Categorical Data.
PUB DATE	1999-01-00
NOTE	17p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Classification; *Coding; Elementary Secondary Education; Inclusive Schools; *Regression (Statistics); Special Education
IDENTIFIERS	*Categorical Data

ABSTRACT

In education and the social sciences, problems of interest to researchers and users of research often involve variables that do not meet the assumptions of regression in the area of an equal interval scale relative to a zero point. Various coding schemes exist that allow the use of regression while still answering the researcher's questions of interest contextually. The coding alternatives, which are used to create "dummy" or "effect" variables, are illustrated using data from a study of special education inclusion (L. Daniel and D. King, 1998). The application illustrates that categorical variables may be successfully combined in regression analyses with continuous variables. For dichotomous predictors, the coding scheme is arbitrary as long as each of the two categories is assigned a numerically different value. Categorical variables with three or more categories can yield varying results depending on the coding scheme used. (Contains 1 table and 10 references.) (SLD)

TM

ED 428 117

FRAGILE--HANDLE WITH CARE: REGRESSION ANALYSES THAT INCLUDE CATEGORICAL DATA

Diane Peacock Brown, MOT, OTL

University of North Texas and Texas Woman's University

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX, January 21-23, 1999.

Linear least-squares regression, and, by extension, weighted regression, nonparametric regression, and general linear models, have become the statistical methods of choice for many researchers (Fox, 1997). There are, however, several assumptions of linear regression, which must be met by the data being analyzed. These include normality, equal variance, and linearity (Fox, 1997).

In education and the social sciences, problems of interest to researchers and consumers of research often involve variables which do not meet the assumptions of regression in the area of an equal interval scale relative to a zero point (Hardy, 1993). For example, categorical variables, such as gender, ethnicity, and intact groups, are often useful variables for consideration in the regression case even though these variables do not fit neatly into the regression model. The researcher must choose among several options, none of which may be particularly desirable: (a) exclude the categorical variables from the analysis; (b) make the variables fit into the analysis in some way; or (c) analyze the data separately for each group within the categorical variable.

The most obvious (and simplest) solution would be to exclude the "problem" variables from the analysis. Unfortunately, theory or reality often dictates that these categorical variables be included in order to accurately measure all factors which may be contributing to the particular phenomenon which is the object of the research. The second solution, making the variables fit into the analysis, has obvious problems when the regression model is used. The researcher may be forced to isolate variables within analyses or else utilize non-parametric techniques, neither of which may be able to answer all the research questions effectively and which may not honor the larger contexts in which particular variables occur. The third solution, analyzing the data separately for each group within the categorical variable, is problematic, particularly if the aim of the research is to compare differences among groups. Given these alternatives, what is the frustrated researcher to do?

Fortunately, there exist several techniques which will allow the use of regression, while still answering the researcher's questions of interest contextually (Cohen, 1968). These consist of various coding schemes, and they are used to create so called "dummy" or "effect" variables, which may then be entered into the analysis. These coding alternatives will be described herein. As with many other areas in which one is faced with choices, there are advantages and disadvantages associated with the different coding methods (Blair & Higgins, 1978). The researcher must choose carefully among them, based upon the particular characteristics of the data in order to avoid sources of error in the analysis.

Sample Data

To illustrate the salient points of this discussion, a portion of an existing data set from a study on special education inclusion described by Daniel and King (1998) was utilized. Daniel and King (1998) studied the effects of inclusion upon four sets of dependent variables was examined, as follows, (a) parent concerns about their children's school programs; (b) teacher and parent-reported instances of students' problem behaviors; (c) students' academic performance; and (d) students' self-reported self-esteem. In the Daniel and King (1998) study, students were divided into three groups, and these categories reflected the method by which children were placed into the classroom groups. Categories were (a) non-inclusion classrooms; (b) "clustered" inclusion classrooms; and (c) "random" inclusion classrooms. This division yielded three intact groups of students in grades 3-5, placed into classrooms by different methods.

For the present, the Stanford Achievement Test (SAT) (The Psychological Corporation, 1990) total score will be used to represent academic achievement, and the Self Esteem Index (SEI) (Brown & Alexander, 1991) total score will represent students' reported self esteem. The variable named special needs reflects the condition (yes or no) of whether or not the child was identified as needing special education services under PL 94-142, the Education for All Handicapped Children Act (1975) and the Individuals with Disabilities Act (IDEA) (1991). Data are used herein for heuristic purposes and may not necessarily represent meaningful analyses per the applied framework presented by Daniel and King (1998).

Dichotomous variables

In the case of a dichotomous variable, the problem of including the categorical variable in the regression analysis is solved by simply assigning a unique value to each of the levels of the variable (Hinkle & Oliver, 1986). The researcher may choose to assign 0 and 1, or may use any other unique values, as will be illustrated in the following example using three coding schemes for the dichotomous variable, special needs.

Three regression analyses were performed, with achievement as the dependent variable and self-esteem and student status (identified special needs or non-special needs) as predictors. The student status variable was coded for each analysis by a different scheme: (a) non-adjusted values, using 1= yes and 2=no (resulting R^{2}= .255, beta weights of -.414 and .230, respectively for special needs and self-esteem); (b) using arbitrary values (i.e., 1999=yes and 666=no) (resulting R^{2}=.255, beta weights of -.414 and .230, respectively); (c) "dummy" coding, using 0=yes and 1=no (resulting R^{2}= .255, beta weights of -.414 and .230, respectively).

(See attached analyses 1a, 1b, and 1c in Table 1). It can be seen that the results are identical, regardless of which two values are used for the two categories within the variable. Dichotomous categorical variables, therefore, can easily be included in a regression analysis so long as each category is assigned a numeric value distinct from the value assigned the other category, and may co-exist with continuous variables.

Polytomous variables

When a variable of interest consists of more than two levels, several coding options exist (Kaufman & Sweet, 1974). The present study examined three alternatives, non-adjusted values, effect coding, and planned comparisons (contrast coding). When using the various coding methods, it is important for the researcher to be aware of exactly what is being compared (Serlin & Levin, 1985). For "dummy" variables, the reference group is usually coded 0. In effect coding, the reference group is coded -1. In planned comparison (contrast coding), the reference group is coded 1 (Hardy, 1993).

To illustrate, consider the variable of classroom membership (i.e., non-inclusion, random inclusion, clustered inclusion) in the Daniel and King (1998) study on inclusion. In the non-adjusted method, the categorical variable is "forced" into a continuous variable (i.e., 1,2, or 3), and the resulting analysis carries the assumption that, somehow, members of group three possess a greater amount of group membership than members of group one. This is ridiculous, of course, but it illustrates the effect of including polytomous categorical variables in regression without recoding. For this analysis, the three groups were coded 1, 2, and 3, respectively. The results, when regression was performed, were $R^2 = .201$ and beta weights=.293 and -.336, respectively, for group membership and self esteem (SEI). (See example 2a in Table 2.)

Next, data were analyzed using the method of "effect coding". In all methods of "dummy" or effect coding, a variable is recoded into one less column than there are levels of the variable (Hinkle & Oliver, 1986). There were three groups; therefore, two columns were required for recoding the group membership variable. In the first column, values were recoded, as follows: group 1=-1, group 2=1, and group 3=0. Thus, the reference group for the effect1 variable (column one) is the non-inclusion group (group 1), since that group is coded -1. The reference group for the effect2 (column two) variable is the clustered inclusion group (group 2). Regression analysis showed R^2=.206. Structure coefficients for effect1 (non-inclusion)=-.576 and effect2 (clustered inclusion).=-.624. (See example 2b in Table 2.)

Using the planned comparison (contrast coding) method of coding, the variable "newgp10" has non-inclusion for a reference group, while "newgp20" has clustered inclusion for reference. Results of regression are R^2=.206, which is the same as that of effect coding. Structure coefficients were as follows: plancom1 (non-inclusion)=.494 and plancom2 (clustered inclusion)=-.322. The continuous variable self-esteem=.724. (See example 2c in Table 2.)

Discussion

As the above examples illustrate, categorical variables may successfully be combined in regression analysis with continuous variables, provided the researcher uses caution when coding the categorical variables. For dichotomous predictors, the coding scheme is arbitrary so long as each of the two categories is assigned a numerically different value. Categorical variables with three or more categories can yield varying results dependent upon the coding scheme employed.

References

Blair, R., & Higgins, J. (1978). Comments on "contrast coding in least squares regression analysis". *American Educational Research Journal, 15*(1), 149-151.

Brown, L., & Alexander, J. (1991). *Self-esteem index examiner's manual.* Austin, TX: PRO-ED.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426-443.

Daniel, L. B., & King, D. A. (1998). The impact of inclusion education on academic achievement, student behavior, student self-esteem, and parental attitudes: A multivariate investigation. *Journal of Educational Research, 91,* 67-80.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks, CA: Sage.

Hardy, M. A. (1993). *Regression with dummy variables.* Newbury Park, CA: Sage.

Hinkle, D. E. & Oliver, J. D. (1986). Regression analysis with dummy variables: Use and interpretation. *Journal of Vocational Education Research, 11*(2), 17-31.

Kaufman, D., & Sweet, R. (1974). Contrast coding in least squares regression analysis. *American Educational Research Journal, 11*(4), 359-377.

The Psychological Corporation. (1990). *Stanford Achievement Test (SAT).* San Antonio, TX: Harcourt Brace & Company.

Serlin, R. C., & Levin, J. R. (1985). Teaching how to derive directly interpretable coding schemes for multiple regression analysis. *Journal of Educational Statistics, 10*(3), 223-238.

Table 1
Regression Results Using Three Coding Methods for Dichotomous Predictors

Analysis 1a
Regression-special needs nonadjusted (with sei)

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SEITOTAL, special needs child[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SAT94TOT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.505[a]	.255	.247	27.08

a. Predictors: (Constant), SEITOTAL, special needs child

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45939.653	2	22969.826	31.313	.000[a]
	Residual	134238.713	183	733.545		
	Total	180178.366	185			

a. Predictors: (Constant), SEITOTAL, special needs child

b. Dependent Variable: SAT94TOT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	23.977	22.873		1.048	.296
	special needs child	-26.033	4.064	-.414	-6.406	.000
	SEITOTAL	.388	.109	.230	3.557	.000

a. Dependent Variable: SAT94TOT

(table continues)

Analysis 1b
Regression-special needs arbitrary coded with SEI

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	weirdscheme, SEITOTAL[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SAT94TOT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.505[a]	.255	.247	27.08

a. Predictors: (Constant), weirdscheme, SEITOTAL

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45939.653	2	22969.826	31.313	.000[a]
	Residual	134238.713	183	733.545		
	Total	180178.366	185			

a. Predictors: (Constant), weirdscheme, SEITOTAL

b. Dependent Variable: SAT94TOT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-41.095	20.960		-1.961	.051
	SEITOTAL	.388	.109	.230	3.557	.000
	weirdscheme	1.953E-02	.003	.414	6.406	.000

a. Dependent Variable: SAT94TOT

(table continues)

Analysis 1c

Regression-special needs dummy coded

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	dichotomous, SEITOTAL[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SAT94TOT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.505[a]	.255	.247	27.08

a. Predictors: (Constant), dichotomous, SEITOTAL

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45939.653	2	22969.826	31.313	.000[a]
	Residual	134238.713	183	733.545		
	Total	180178.366	185			

a. Predictors: (Constant), dichotomous, SEITOTAL

b. Dependent Variable: SAT94TOT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.056	21.560		-.095	.924
	SEITOTAL	.388	.109	.230	3.557	.000
	dichotomous	-26.033	4.064	-.414	-6.406	.000

a. Dependent Variable: SAT94TOT

Table 2
Regression Results Using Three Coding Methods for Polytomous Predictor

Analysis 2a
Regression - Group (classroom) membership nonadjusted

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SEITOTAL, EXPGROUP[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SAT94TOT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.448[a]	.201	.192	28.05

a. Predictors: (Constant), SEITOTAL, EXPGROUP

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36166.941	2	18083.470	22.979	.000[a]
	Residual	144011.425	183	786.948		
	Total	180178.366	185			

a. Predictors: (Constant), SEITOTAL, EXPGROUP

b. Dependent Variable: SAT94TOT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-8.723	22.295		-.391	.696
	EXPGROUP	-11.674	2.297	-.336	-5.082	.000
	SEITOTAL	.495	.111	.293	4.438	.000

a. Dependent Variable: SAT94TOT

(table continues)

Analysis 2b
Regression-experimental condition, effect coded

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	effectcod2, SEITOTAL, effectcod1[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SAT94TOT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.454[a]	.206	.193	28.03

a. Predictors: (Constant), effectcod2, SEITOTAL, effectcod1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37195.955	3	12398.652	15.782	.000[a]
	Residual	142982.410	182	785.618		
	Total	180178.366	185			

a. Predictors: (Constant), effectcod2, SEITOTAL, effectcod1

b. Dependent Variable: SAT94TOT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-32.668	21.682		-1.507	.134
	SEITOTAL	.493	.111	.292	4.424	.000
	effectcod1	-9.072	3.051	-.202	-2.974	.003
	effectcod2	-4.903	1.408	-.236	-3.482	.001

a. Dependent Variable: SAT94TOT

(table continues)

Analysis 2b continued

Correlations

		effectcod1	effectcod2	Unstandardized Predicted Value
effectcod1	Pearson Correlation	1.000	.230**	-.576**
	Sig. (2-tailed)	.	.002	.000
	N	186	186	186
effectcod2	Pearson Correlation	.230**	1.000	-.624**
	Sig. (2-tailed)	.002	.	.000
	N	186	186	186
Unstandardized Predicted Value	Pearson Correlation	-.576**	-.624**	1.000
	Sig. (2-tailed)	.000	.000	.
	N	186	186	186

**. Correlation is significant at the 0.01 level (2-tailed).

Analysis 2c
Regression-experimental condition, planned comparison

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	plancom2, SEITOTAL, plancom1[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SAT94TOT

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.454[a]	.206	.193	28.03

a. Predictors: (Constant), plancom2, SEITOTAL, plancom1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37195.955	3	12398.652	15.782	.000[a]
	Residual	142982.410	182	785.618		
	Total	180178.366	185			

a. Predictors: (Constant), plancom2, SEITOTAL, plancom1

b. Dependent Variable: SAT94TOT

(table continues)

Analysis 2c continued

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-42.474	21.765		-1.952	.053
	SEITOTAL	.493	.111	.292	4.424	.000
	plancom1	23.781	4.606	.360	5.163	.000
	plancom2	5.637	5.753	.068	.980	.328

a. Dependent Variable: SAT94TOT

Correlations

		Unstandardized Predicted Value	plancom2	plancom1	SEITOTAL
Unstandardized Predicted Value	Pearson Correlation	1.000	.301**	.494**	.724**
	Sig. (2-tailed)	.	.000	.000	.000
	N	186	186	186	186
plancom2	Pearson Correlation	.301**	1.000	-.322**	-.014
	Sig. (2-tailed)	.000	.	.000	.845
	N	186	186	186	186
plancom1	Pearson Correlation	.494**	-.322**	1.000	.015
	Sig. (2-tailed)	.000	.000	.	.841
	N	186	186	186	186
SEITOTAL	Pearson Correlation	.724**	-.014	.015	1.000
	Sig. (2-tailed)	.000	.845	.841	.
	N	186	186	186	186

**. Correlation is significant at the 0.01 level (2-tailed).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM029588

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: "Fragile--Handle with Care: Regression Analyses That Include Categorical Data."	
Author(s): Diane P. Brown, MOT, OTL	
Corporate Source: Southwest Educational Research Assoc. Annual Meeting, 1999.	Publication Date: January 21-23, 1999

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Diane P. Brown	Printed Name/Position/Title: Diane P. Brown, MOT, OTR	
Organization/Address: Texas Woman's University School of Occupational Therapy P.O. Box 425648 Denton, TX 76204	Telephone: 940.898.2812	FAX: 940.898.2806
	E-Mail Address: f_2brown@twu.edu	Date: 012799

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com